The Capability Firewall

Once intelligence can act, capability becomes a security boundary. The question was never which model. It is what that capability has been wired to, and whether anything enforces the line.

The frontier labs have started grading their own models by capability, drawing explicit thresholds for how much cyber reach and autonomy is too much to release without new controls. Most people read that as a safety milestone. It is also an admission. The danger they are measuring is not intelligence. It is capability joined to reach. And the question it forces on every institution deploying AI is no longer which model is approved. It is what that model has been connected to, and what holds the line when it acts.

AUTHOR
Shane Schreck

PUBLISHED
September 2026

CANONICAL URL
https://publications.aleeth.com/the-capability-firewall/

Publication Record

Publication	Essay
Published	September 2026
Author	Shane Schreck
Canonical URL	https://publications.aleeth.com/the-capability-firewall/

The labs are now publishing capability thresholds. A model that can find unknown vulnerabilities, write working exploits, and pursue a goal across many steps with little human guidance is no longer described as merely capable. It is described as a threshold to be gated. The people who build these systems are telling the market, in their own preparedness documents, that past a certain point capability itself is the risk.

They are right. And they have named the exact place governance has to move.

SAME MODEL, FOUR ROOMS

Take one capable model and put it in four settings, unchanged.

In the first, it has no tools and only public information. It is an advisor. In the second, it can read the enterprise and retrieve from it. It is a knowledge worker. In the third, it can call internal systems and move a workflow. It is an operator. In the fourth, it can execute code, reach the network, and hold production credentials. It is an autonomous agent with real destructive reach.

Same intelligence in every room. The consequence is not the same, and it is not close. The risk did not come from the model. It came from what the model was wired to.

Capability is not dangerous until it is connected. The connection is the boundary. A model safe enough to summarize a document can be unacceptable the moment it is given a key to production, and nothing about the model changed.

RISK IS A PRODUCT, NOT A PROPERTY

This is why the question so many organizations still ask, which model are we allowing, which app can employees use, which vendor is approved, is going quietly obsolete. Those questions treat risk as a property of the model. Risk is not a property. It is a product.

Capability, times context, times access, times authority, times autonomy. No single factor is the danger. The danger emerges from how they combine.

And because it emerges from the combination, it has to be judged on the action, not on the model, and judged every time. At least eight dimensions decide it: how much autonomy the action carries, what tools it can reach, what privilege it runs with, how far its network reach extends, how sensitive the data it touches is, whether it leaves persistent state, whether it can be reversed, and how large its blast radius is if it is wrong.

Role based access control cannot carry this. It was built for a human making one decision at a login screen, granted a role and trusted with it. An autonomous system makes authorization a continuous problem, thousands of actions a day, each with a different combination behind it. The permission model built for one person at a door cannot govern a system that decides and acts in the same motion.

A FIREWALL DECIDES, OR IT IS JUST ADVICE

A network firewall earns its name because it does not counsel the packet. It permits or it drops. Anything less is a suggestion. A capability firewall has to hold the same standard for an agent's actions.

For every governed action it returns one verdict, enforced, denying by default. Read, permitted. Recommend, permitted. Execute, only inside defined boundaries. Human approval, required before it proceeds. Never, this capability does not reach this environment at all.

Not advice the agent may take or ignore. Enforcement. And it fails closed. If the decision cannot be governed and sealed, the action does not happen. No verdict, no execution. That single property, deny by default and fail closed, is the line between a control plane and a dashboard that watches harm scroll past.

THE RECEIPT IS THE PROOF

This is where most of what is sold as AI governance stops, at the edge of theater. Detection is not attestation. A screen that says a system is compliant is a claim. A claim is not proof.

So every governed decision has to leave a receipt, and the receipt has to be one no one can quietly rewrite. In this architecture each decision is cryptographically signed and hash chained into an append only ledger, and any receipt can be verified in a browser, offline, with no call back to us and no trust in us required. As of today that ledger holds more than 209,000 signed decision receipts, and it grows with every governed action.

The record that proves the action was controlled is the same record that proves what the system actually did. Governing the act and proving the act are not two jobs. They are one.

THE ONE THING A LAB CANNOT BE

There is a role the model companies cannot fill, no matter how good their intentions. A company cannot be the neutral referee of its own capability. The control and assurance layer has to be independent of the model vendor and has to govern any model, including the frontier ones, including the next one no one has shipped yet.

That independence is not a feature. It is the whole point. An assurance layer that belongs to the thing it is assuring is a marketing department. The layer that decides what an agent may do, and seals the proof that it did only that, has to answer to the institution deploying the AI, not to the lab that built it.

Beneath the act. Above the machine. Impossible to bypass, across the surfaces it governs.

Capability is not dangerous until it is connected. The connection is the boundary, and a boundary that is not enforced is not a boundary at all.

The labs have told us where the line is. They are grading their own models against it. What they cannot do is stand on both sides of it. Every institution now running autonomous AI has one question to answer for every action that AI takes: can you prove it stayed inside the authority you granted. The only real answer is a firewall for

capability, enforcing the line at runtime and sealing the proof.

The full standard, Institutional Control Architecture, is published at publications.aleeth.com/standard . The instrument is in production today.

Not pitched. Not promised. Proven.

Shane Schreck is the founder of ALEETH and the author of Institutional Control Architecture. ALEETH is building the independent verification layer for autonomous AI deployment. He is a U.S. Army Veteran.

ALEETH

AI has met its match. Truth.™